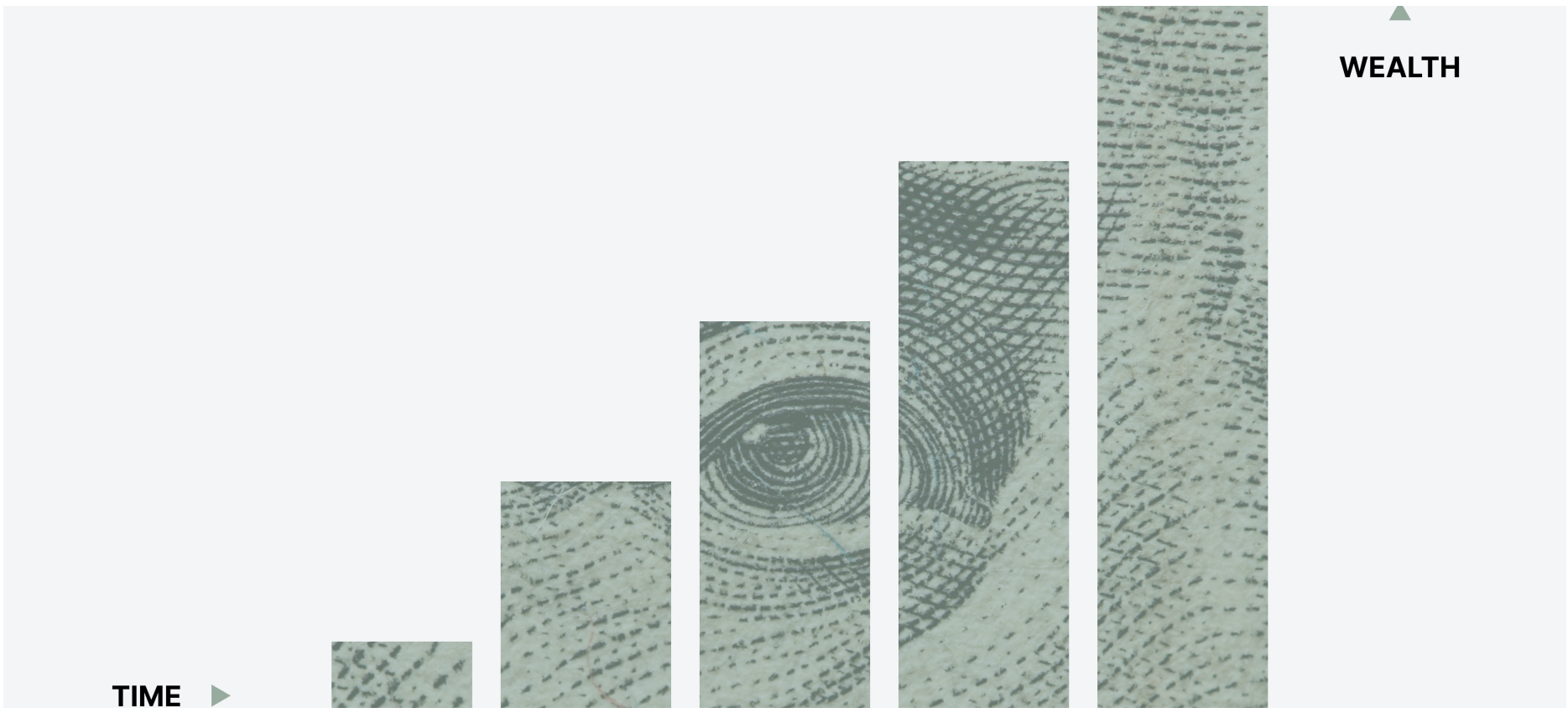

Week 4 : Descriptive Statistics (Part 1)

pending tasks

type

Introduction to descriptive statistics

The learning objectives are to answer the following questions:

1. What are different types of data?
2. How do we describe qualitative data?
3. How do we describe quantitative data?
4. How do we describe relationships between attributes?

Different types of data

def qualitative data : (t = 2:25)

def nominal data : (t = 3:15)

def ordinal data : (t = 4:16)

def quantitative data : (t =9:33)

def discrete data : (t = 10:05)

def continuous data : (t = 10:41)

The type of statistical analysis used depends on the data type. the data can be divided into the following types:

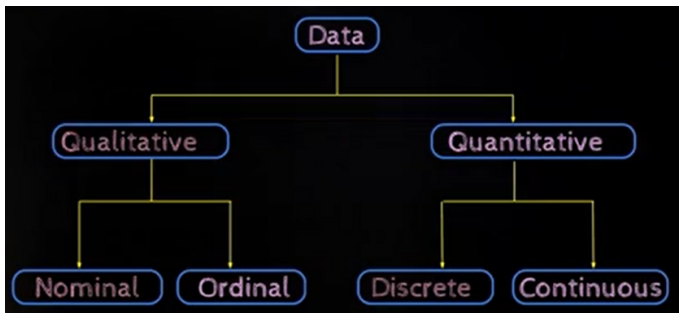

Fig.1 Types of data.

Qualitative data : categorical attributes which describe the object under consideration using a finite set of discrete classes. Eg.colour, rating and size. Statistical measures such as frequency, analysis of variance (ANOVA),chi-square test can be done and frequency charts can be plotted.

Nominal data : qualitative data with no natural ordering. Eg. colour of the cloth. Even if the categories are mapped to numbers, the notion of distance is not sell defined.

Ordinal data : qualitative data with natural ordering. Eg. customer rating.

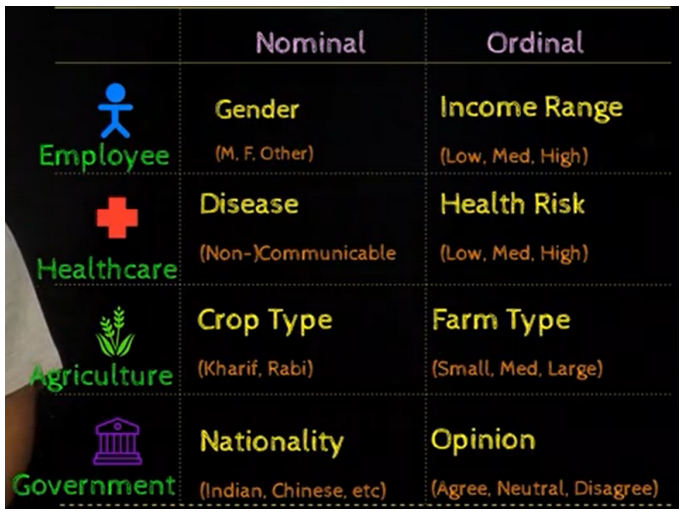

Fig.2 Examples of nominal and ordinal data.

Quantitative data : attributes with numerical values which are used to measure certain properties of the population. Statistical measures such as mean, median, mode can be calculated and histograms can be plotted. Regression analysis can be done.

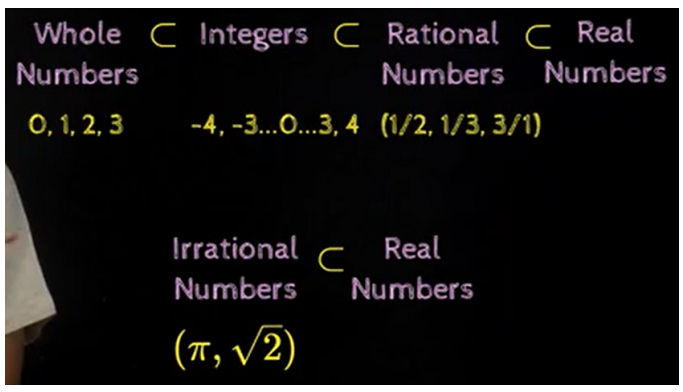

Fig.3 types of numbers.

Discrete data : quantitative attributes that can take up only integers.

Continuous data : quantitative attributes which can take up fractional values also.

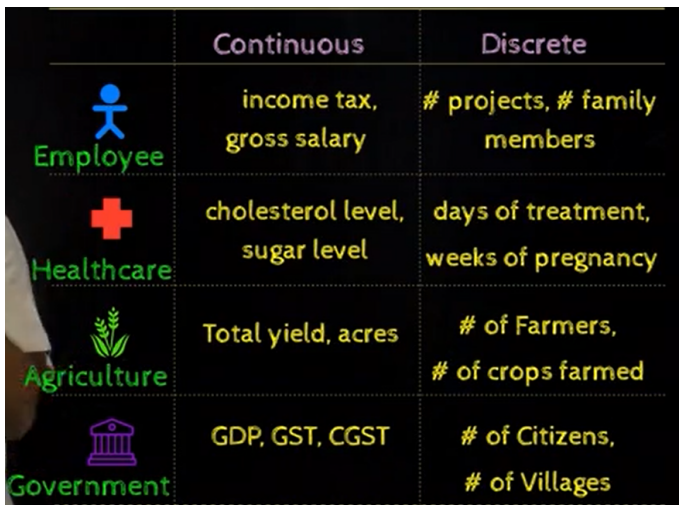

Fig.4 Examples of continuous and discrete data.

How to describe qualitative data?

def frequency - (t = 1:40)

- The values of categorical data types keep repeating in the data. Thus, frequency of occurrence can be found and meaningful insights can be derived using grouping of attributes. Comparative analysis of frequencies of different categories can be done.
- For a quick analysis of categorical attributes, **frequency plots**, derived from frequency tables can be used. Here sorting the plots based on frequency is a convention used for easier interpretation.
- Data distribution can be long tailed or uniform.

- **Relative frequency plots** are used to answer questions about percentages. Relative frequency is the frequency/total number of frequencies. It is easier to interpret relative frequencies than absolute frequencies.
- To compare different sets of data **grouped frequency bar charts** are used. Here using relative frequencies makes more sense. Thus, **grouped (rel.) frequency bar charts** are used.

Frequency plots are used in ML to analyse the input data categories, output error, designing features.

How to describe quantitative data?

def left-end-inclusion-convention - (t = 32:14)

Discrete Data

- Frequency of different values can be found using a **histogram**. Here the values on the x-axis are numbers instead of categories and the values are sorted by natural order.
- The x-axis values can be binned into class-intervals. This is done based on domain requirement. As the bin size is increased, the granularity is compromised. It depends on the range of the data.
- Ideal bin size reveals meaningful patterns (neither hides nor reveals too many details).

Continuous data

- The values have to be binned to reveal meaningful patterns in the data.
- Left-end-inclusion convention is used → **[ ).**

Steps to plot a histogram :

1. Sort the values in increasing order.
2. Choose class intervals such that it covers all the values.
3. Compute frequency of each interval.
4. Draw bars for each interval.

Histograms

- To answer the questions on percentages and comparison, relative frequencies are found for respective class intervals.

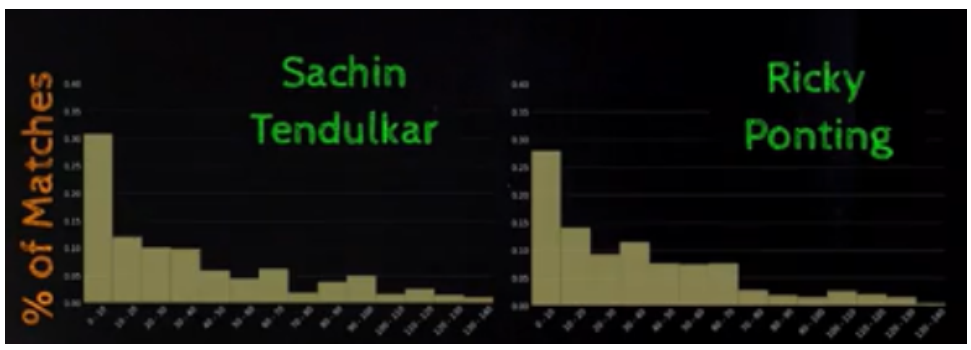

Fig.5 Relative score comparison, x-axis is class intervals of runs.

Steps to plot Rel. Freq. histograms :

1. Sort the values in increasing order.
2. Choose class intervals such that it covers all the values.
3. Compute relative frequency of each interval.
4. Draw bars for each interval.

- When comparison is done among, eg. many players,all the histograms can be drawn in one plot. But this makes it hard to distinguish between individual histograms. Instead, Grouped bar charts can be drawn as in Fig.7.

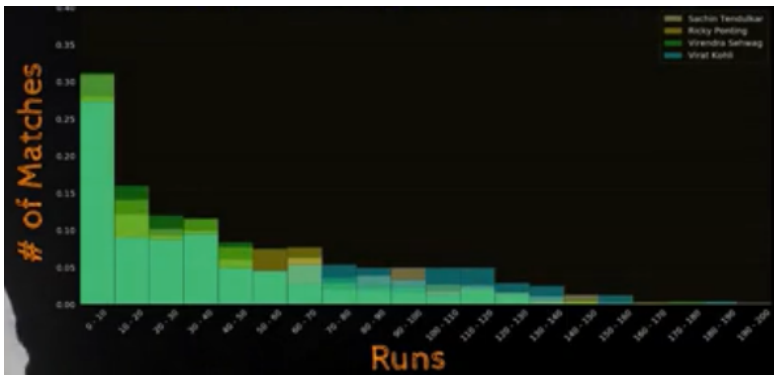

Fig.6 Drawing all histograms in one plot.

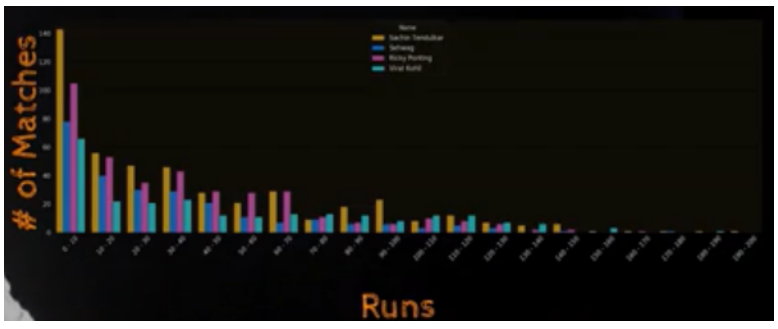

Fig.7 Grouped histogram.

- **Frequency polygons** can also be used. This helps distinguishing and comparing overall trends for different players easier.

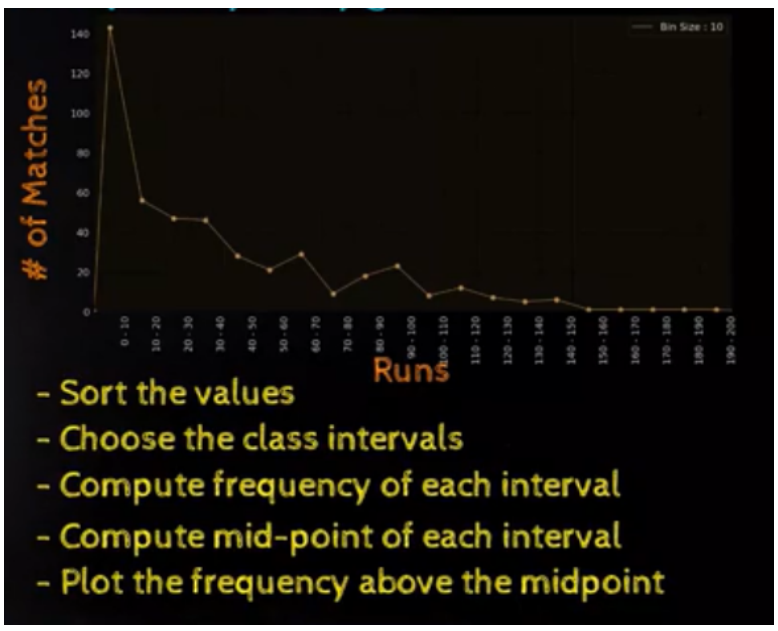

Fig.8 Frequency polygon.

Here , drawing a relative **frequency polygon gives the right scale for comparison** of data.

Cumulative frequency polygon

- For each class interval, the sum of frequencies of all preceding class intervals are added. This can be plotted for relative frequencies, (**cumulative relative frequency polygon**) to answer the questions involving percentages and comparison of sets of data.

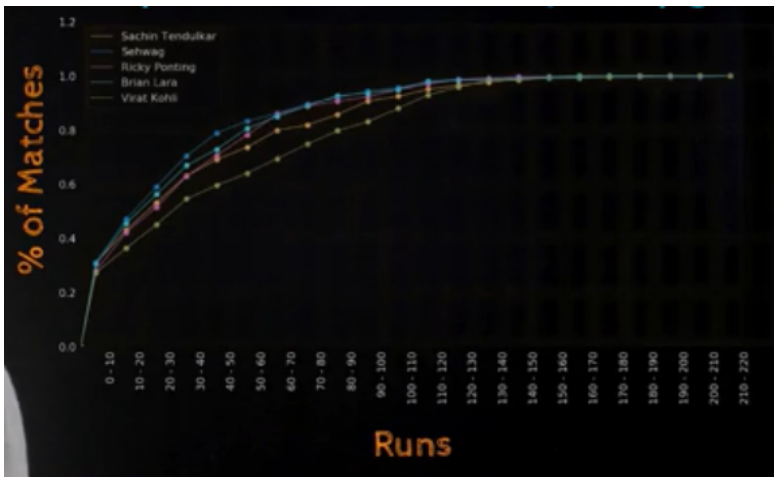

Fig.9 Cumulative relative frequency polygon.

Typical trends in histograms

1. Spread of data
2. Data density of intervals.
3. Gaps in the data.
4. Outliers in the data.

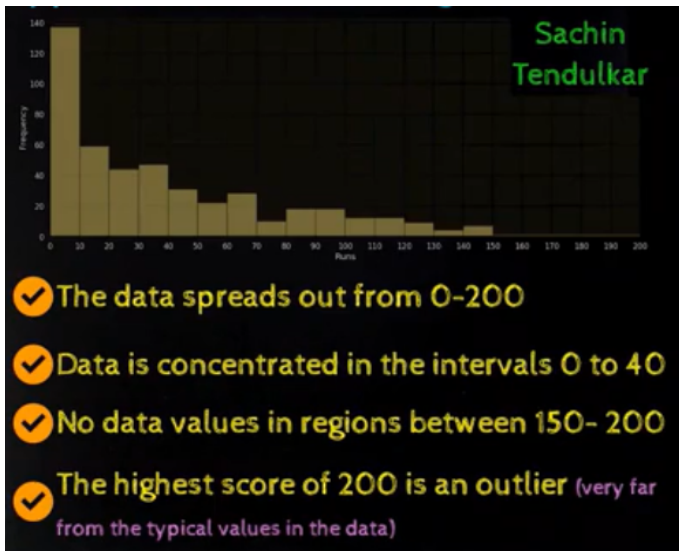

Fig.10 Example of trends in histogram.

- Histograms can be left/right skewed (direction of the tail) and can have heavy left/right tail.
- In uniform histograms, most bars are of similar height.
- In symmetric histograms, the bars are almost mirrored images of each other about the vertical median.

Uses of histograms in ML

1. Identifying discriminatory features.
2. Analysing output scores.

Stem and leaf plots

- Efficient way of describing small to medium data, not for large data.
- It represents every number in the data in two parts : stem and leaf.
- For continuous data , the fractional numbers are rounded to the nearest integer and then plotted.
- The stem length has to be chosen so as to display meaningful patterns in the data. The scale of the leaf can be chosen based on the granularity required.
- If there are too many values in a row , it can be divided into two based on leaf values.
- Displaying individual values makes it easy to spot patterns.
- Back to back stem and leaf plots can be drawn to compare two different datasets.

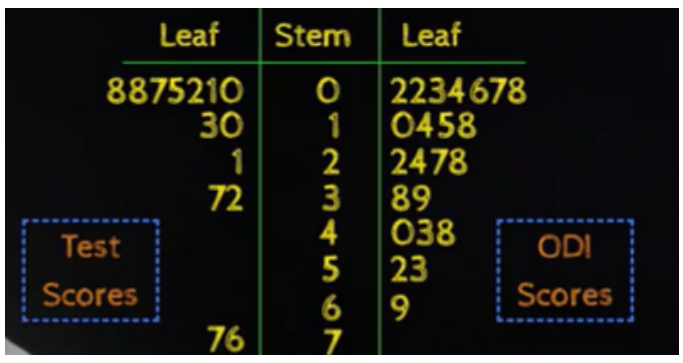

Fig.11 Example of back to back stem and leaf plot.

How to describe relationship between variables? Scatter plots

(for quantitative attributes)

- The correlation between 2 attributes can be found using scatter plots. Patterns not visible in individual histograms can be found in the scatter plots.

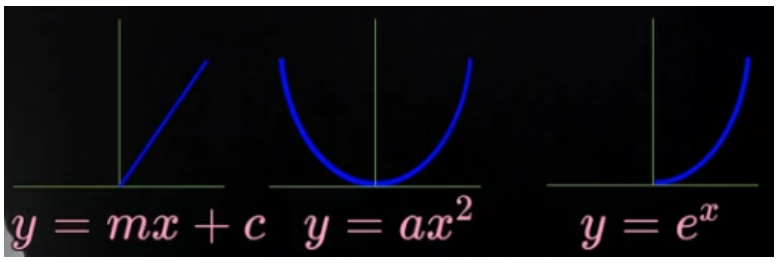

Fig.12 Recap of functions.

- It can be inferred from a scatter plot if two variables are related, unrelated videos imply that the information in one variable cannot be obtained from the other.

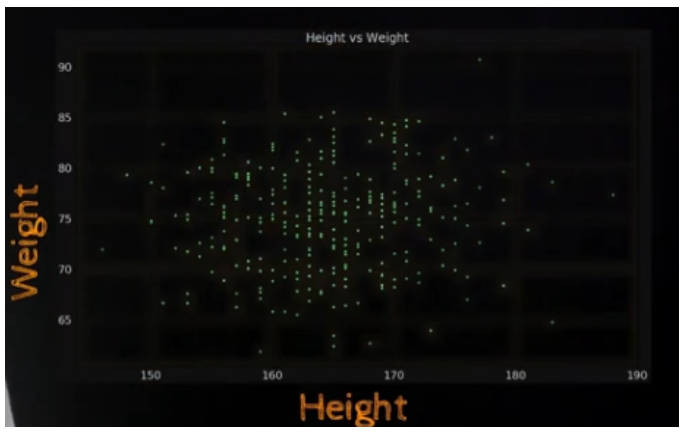

Fig.13 Example of uncorrelated variables.

In ML scatter plots can be used to Identify correlated features . For a model, uncorrelated or non-redundant features are to be used for classification.

Summary

- The data can be broadly classified into quantitative (nominal and ordinal) and qualitative (discrete and continuous).
- To describe qualitative data frequency (relative and absolute) tables, bar charts and grouped bar charts are used.
- To describe quantitative data (relative/absolute) histograms and frequency polygons, stem & leaf plots and scatter plots are used.
- Relative values help in finding percentage values and comparison of data.
- Class intervals are used such that meaningful patterns are easily found in the data.

MCQ : Week 4

1. A student record has name, age,subject, marks secured and rank. What type of data is the student's name, age and marks and rank? Choose the most relevant option.
    1. **Nominal, discrete, continuous, ordinal**
    2. Ordinal, continuous,discrete, ordinal
    3. Qualitative, discrete, continuous, nominal
    4. All are qualitative since the variables define the attributes of the student.
2. The scatter plot of 2 uncorrelated variables
    1. Is a linear relationship.
    2. Is a line with slope 1.
    3. **random plot with no pattern to draw a relationship.**
    4. Initially a linear relationship is portrayed then an exponential one.
3. Relative frequency plots are used to
    1. answer questions involving percentages for quantitative attributes.

2. **answer questions involving percentages for qualitative attributes.**
3. **useful for comparison of a given attribute of different objects.**
4. useful for comparison of different attributes of a given data object.

Note : example of a data object can be a player such as sachin and attributes can be runs scores, balls faced etc.

4. To compare the quantitative attributes of different data objects,
    1. a relative frequency plot can be used.
    2. Independent histograms can be drawn representing the data of each object.
    3. **A relative frequency polygon can be used.**
    4. **A cumulative frequency polygon is used.**
5. A stem and leaf plot
    1. can be used for large datasets if all the outliers and null values are removed.
    2. **is used for plotting fractional numbers by rounding off to the nearest integer values.**
    3. **makes it easy to spot patterns in data as individual values are plotted.**
    4. **can have scaled down leaves to improve readability.**